

Examining Two Strategies to Link Mixed-Format Tests Using Multiple-Choice Anchors

Michael E. Walker

Sooyeon Kim

July 2010

ETS RR-10-18



Linking Mixed-Format Tests Using Multiple-Choice Anchors

Michael E. Walker and Sooyeon Kim
ETS, Princeton, New Jersey

July 2010

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

<http://www.ets.org/research/contact.html>

Technical Review Editor: Dan Eignor

Technical Reviewers: Mary Grant and Gautam Puhan

Copyright © 2010 by Educational Testing Service. All rights reserved.

ETS, the ETS logo, and LISTENING. LEARNING.
LEADING. are registered trademarks of Educational Testing
Service (ETS).



Abstract

This study examined the use of an all multiple-choice (MC) anchor for linking mixed format tests containing both MC and constructed-response (CR) items, in a nonequivalent groups design. An MC-only anchor could effectively link two such test forms if either (a) the MC and CR portions of the test measured the same construct, so that the MC anchor adequately represented the entire test, or (b) the relationship between the MC portion and the total test remained constant across the new and reference linking groups. The study also evaluated whether linking mixed-format tests through MC-only anchors would be more effective than a two-stage strategy in which MC portions were equated through MC anchors and then composite scores were scaled to the MC scores. Anchor linking and two-stage linking yielded identical (or nearly so) results for both linear and nonlinear chained linking methods. With post-stratification linking methods the two-stage strategy resulted in smaller bias. The paper discusses some advantages of both approaches.

Key words: mixed-format test, constructed response, equating, scaling, NEAT design

Acknowledgments

We would like to thank Mary C. Grant, Gautam Puhan, and Daniel Eignor for their many helpful comments on the earlier draft of this paper. The authors also gratefully acknowledge the editorial assistance of Kim Fryer.

Table of Contents

	Page
A Linking Criterion.....	3
The Hypotheses.....	5
Method	6
Data.....	6
Simulated Forms.....	6
Procedure	7
Results.....	10
Descriptive Statistics	10
Linking Results.....	11
Bootstrapping Results.....	14
Regression Results.....	18
Discussion.....	19
References.....	23
Notes	25

List of Tables

	Page
Table 1 Symbols Used to Represent Composite and Multiple Choice (MC) Scores for Reference and New Forms in Illustrating Equivalence of Linking Criteria for Direct and Two-Stage Linking.....	5
Table 2 Means (Standard Deviations) and Correlations for Examinee Groups Taking New and Reference Forms	10
Table 3 Summary of Root Mean Squared Difference (RMSD) Between Linking Results and Criterion for Two Linking Strategies with Chained and Post-Stratification Linking Methods.....	12
Table 4 Summary of Bootstrapped Weighted Average Root Mean Squared Bias, Equating Error, and Root Mean Squared Error (RMSE) for the Two Linking Strategies and Two Methods.....	14
Table 5 Summary of Bootstrapped Weighted Average Root Mean Squared Bias, Equating Error, and Root Mean Squared Error (RMSE) for Post-Stratification Methods.....	16

List of Figures

	Page
Figure 1. Three strategies for linking the new composite to the reference composite using multiple-choice (MC)-only anchors.....	3
Figure 2. Schematic of the criterion and linking designs examined in this study.	7
Figure 3. Relative frequency polygon for the new form composite score in the new form group.	11
Figure 4. Difference between chained linking and the criterion for the direct anchor and two-stage strategies, chained linear and chained equipercentile linking methods.....	12
Figure 5. Difference between post-stratification linking and the criterion for the direct anchor and two-stage strategies, Tucker and frequency estimation equipercentile linking methods.....	13
Figure 6. Conditional bias (shown as the average difference of linking results from the criterion) for the direct anchor and two-stage strategies, chained linear case). Linking error bands are also shown.	15
Figure 7. Conditional bias (shown as the average difference of linking results from the criterion for the direct anchor and two-stage strategies, chained equipercentile case). Linking error bands are also shown.	15
Figure 8. Conditional bias (shown as the average difference of linking results from the criterion for the direct anchor and two-stage strategies, Tucker case). Linking error bands are also shown.	17
Figure 9. Conditional bias (shown as the average difference of linking results from the criterion for the direct anchor and two-stage strategies, frequency estimation equipercentile case). Linking error bands are also shown.	17
Figure 10. Cubic regression functions relating composite score to total multiple-choice score for the reference and new groups.....	18

Many large-scale testing programs include constructed-response (CR) as well as multiple-choice (MC) items in their assessments. As with other standardized tests, these mixed-format tests must be equated to ensure equivalence of scores across test forms. Perhaps most often, equating occurs in the context of the nonequivalent groups with anchor test (NEAT) design, in which a set of items common to both the new and reference forms is used to place scores on both forms on the same scale. These common items should represent the entire test form in terms of content and difficulty.

NEAT equating has proven difficult with mixed-format tests. One reason is that because CR items tend to be easy to memorize (Muraki, Hombo, & Lee, 2000), it may be difficult to find CR items that can be reused across forms. Second, even if the same CR items are used, the standards of the raters scoring the CR items almost always differ across the two administrations (Fitzpatrick, Ercikan, Yen, & Ferrara, 1998). In this case the CR anchor items would confound differences in rater severity with true group differences, so that adjustment through the anchor could lead to erroneous results (Tate, 1999).

One solution to the second problem involves rescoring CR papers from the reference form administration at the new form scoring (a process called trend scoring). For a random group of examinees from the previous administration, the anchor CR items are rescored by the raters who score the new form in the current administration. The scores from the new raters (not the old) on the reference form become part of the anchor score. This method effectively nullifies any differences in scoring standards across administrations, because the same raters score the anchor items on both sets of papers. Equating with trend scoring has proven successful in producing accurate linkings (Kim, Walker, & McHale, 2008, 2010; Tate, 1999, 2000). However, the time, logistic, and monetary costs of this method could be prohibitive in certain situations.

Another possible solution to the problem of lack of an appropriate CR anchor would be to attempt to link mixed-format tests through MC items only. Indeed, some practitioners have suggested using MC items as anchors to control for differences among test forms containing CR items (e.g., Baghi, Bent, DeLain, & Hennings, 1995; Ercikan et al., 1998). However, evidence suggests that using an all-MC anchor can lead to biased equating results (Kim & Kolen, 2006; Kim et al., 2008, 2010; Li, Lissitz, & Yang, 1999), possibly because MC and CR items measure somewhat different constructs (Bennett, Rock, & Wang, 1991; Sykes, Hou, Hanson, & Wang, 2002).

This research focuses on the use of an internal (i.e., part of the total test) MC-only anchor to place two mixed-format test forms on the same scale. There are at least two different ways we can envision linking through MC-only anchors. The first possibility is that if the MC part of the test measures the same construct as the CR part, and if the CR portions of the test are fairly consistent from form to form with respect to content and difficulty, then adjusting for group ability through an MC-only anchor would result in what could be called equating, in the sense that the composite score for the new test would be aligned along the same dimension as the composite score for the reference test.

A second possibility is that the MC part of the test measures a somewhat different construct from the total test, or that the CR portions vary from form to form with respect to content and rater standards. In the context of the NEAT design, we could use scale alignment techniques to place the reference and new test scores on a common scale (Holland & Dorans, 2006). To do this we could equate the MC portions of the reference and new test forms using the MC anchor. We could then scale both composite (MC plus CR) scores to the MC scores, so that both composite scores would be on the common MC score metric. Note that once both MC portions have been equated, the MC scores on the two tests may be considered equivalent. Thus, linking the composite scores in this way is an example of what Holland and Dorans (2006) called scaling to an anchor. Even if the MC and CR items measured somewhat different constructs, this method would still result in comparable scores, as long as the relationship between the MC and CR sections were identical across the reference and new examinee populations, leading to constant dimensionality across populations. This research examined whether either of these conditions (unidimensionality vs. constant dimensionality) held, such that mixed-format tests could successfully be linked through MC-only anchors.

Another purpose of this study was to evaluate a claim (M. Kolen, personal communication, November 26, 2007) that linking mixed-format tests through MC-only anchors would result in something closer to equating than would a strategy in which MC portions were equated through MC anchors and then composite scores were scaled to the MC scores. One could counterclaim that the MC portions of the tests are properly equated using the MC anchor, whereas the composite scores are not. Furthermore, the relationship between the MC portion and the composite should be stronger than that between the MC anchor and the composite by virtue of the longer length of the entire MC section. Thus we might expect scaling to an anchor to align the two test scores better

than linking through the MC-only anchor. This advantage would be seen in the case of post-stratification methods, which make explicit use of bivariate information.

A Linking Criterion

To better understand the relationship between anchor linking and scaling to an anchor in this context, it is worthwhile to investigate what the linking criteria should be for the two strategies. In the more direct anchor linking, the new composite is placed on the scale of the reference composite via the MC anchor tests. Figure 1A illustrates this procedure. In the figure, all arrows lead from the new composite to the reference composite. The goal of linking here is to estimate what the new-to-reference score relationship is in the population. Theoretically, the criterion would be found by directly linking the new scores to the reference scores in a population in which every individual had taken both tests.

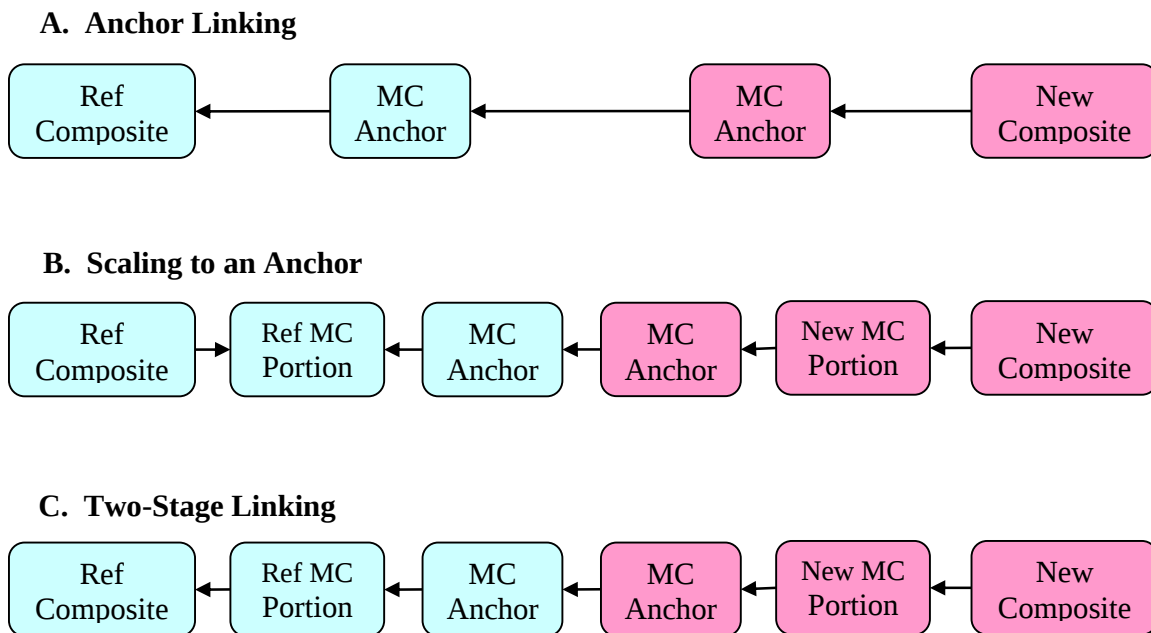


Figure 1. Three strategies for linking the new composite to the reference composite using multiple-choice (MC)-only anchors.

With scaling to an anchor, illustrated in Figure 1B, the new MC portion is placed on the scale of the reference MC portion. Then both composites are placed on this reference MC scale through direct scaling. In the figure, all arrows lead to the reference MC portion. The criterion for such a linking would be on the scale of the reference MC portion.

Note that the scales in Figures 1A and 1B are different. While 1A places scores on the scale of the reference form composite score, 1B places scores on the scale of the reference form MC score. To compare the two strategies and their criteria directly, we would need to place the scores resulting from each strategy on the same scale, say the scale of the reference composite score. This is easily accomplished by reversing the arrow in Figure 1B leading from the reference composite to the reference MC. This procedure yields Figure 1C.

We argue that this reversal has not changed the essential nature of the relationships among the various scores because the methods used in equating and scaling are symmetric by nature (see Holland & Dorans, 2006; Kolen & Brennan, 2004). This means that the function mapping one set of scores onto the other is one-to-one and therefore invertible. Thus we can take the arrow leading from the reference composite to the reference MC portion in Figure 1B and reverse it to get Figure 1C. The relationships implied by Figures 1B and 1C are equivalent, but they are on different scales. The scale in Figure 1C is that of the reference composite score, the same as in Figure 1A. As in Figure 1A, all arrows in Figure 1C lead from the new composite to the reference composite. The figure would imply that the criterion for Figure 1C is the same as for Figure 1A and is found by directly linking the new composite scores to the reference composite scores in a population in which every individual has both scores. Thus Figure 1C represents a two-stage linking strategy whose goal is equivalent to the more direct anchor linking method depicted in 1A.

We can illustrate this point algebraically as well by using the equipercentile linking function. To do so, we adopt the notation of Braun and Holland (1982). Table 1 shows the symbols we will use for the scores involved in forming the criteria: the variable name and the associated cumulative distribution function (cdf) for each. Note that the MC-only anchor tests are not included in the table, because these would theoretically not be needed when forming the criterion in the population.

Table 1

Symbols Used to Represent Composite and Multiple Choice (MC) Scores for Reference and New Forms in Illustrating Equivalence of Linking Criteria for Direct and Two-Stage Linking

Score	Symbol	cdf
Ref composite	Y	$G(y)$
Ref MC	V	$H(v)$
New MC	W	$E(w)$
New composite	X	$F(x)$

Note. cdf = cumulative distribution function.

Assume that the scores in Table 1 are continuous random variables. Then, using the symbols from the table, the direct equipercentile function linking new composite (X) scores to reference composite (Y) scores may be written (Braun & Holland, 1982) as

$$e_Y(x) = G^{-1}(F(x)). \quad (1)$$

In words, the function that places new score X on the scale of old score Y is a compound function involving the inverse cdf of Y and the cdf of X . Likewise, the two-stage linking strategy may be expressed as

$$e_Y(x) = G^{-1}\left(H\left\{H^{-1}\left(E\left[E^{-1}(F(x))\right]\right)\right\}\right). \quad (2)$$

Cancellation of functions with their inverses leaves a result identical to Equation 1. In other words, both the direct anchor and two-stage strategies yield the same criterion in this case.

The Hypotheses

The current research examined both direct anchor linking and a two-stage process to assess the effectiveness of an MC-only anchor in linking mixed format tests. Both linear and nonlinear linking methods were used. Both chained (linear and equipercentile) and post-stratification

(Tucker and frequency estimation equipercntile) linking methods were examined. We hypothesized the following: First, that the use of an MC-only anchor would effectively link two tests to the extent that the MC-composite relationship remained constant across the reference and new form groups; second, that the direct anchor and two-stage strategies would yield identical results in the chained linking case; finally, that the two-stage strategy would yield superior results (i.e., closer to the criterion) in the post-stratification linking case.

Method

Data

The data for the study were taken from two administrations of a subject test, comprising 24 MC and 12 CR items, of a large-scale testing program. Each MC item received a maximum score of 1, and each CR item received a maximum score of 4. Thus, the possible score range of this test (called Form Z) was 0 to 72.

For one administration, the 12 CR items for 417 examinees were scored by Rater Group A. These 417 examinees constituted the reference group in this study. In another administration, the same 12 CR items for those 417 examinees were independently scored by Rater Group B. These same Raters (Group B) also scored the 12 CR items for a separate group of examinees ($N = 3,126$). These 3,126 examinees constituted the new group in this study. Note that two independent sets of scores for all CR items were available for the 417 reference examinees, but only a single set of CR scores was available for the 3,126 new examinees.

Simulated Forms

The original test form used in this study had 24 MC and 12 CR items. Two forms parallel in both content and difficulty (designated new form and reference form) were created from the original form. The new and reference forms consisted of 16 MC and 8 CR items. Those forms had 8 MC and 4 CR items in common. Only the common MC items were used as the anchor in a NEAT design. The maximum possible scores for the test and anchor were 48 and 8, respectively. For the purposes of the study, the reference form as scored by Rater Group A (reference form/Rater A) served as the reference form, and the new form as scored by Rater Group B (new form/Rater B) served as the new form.¹

The construction of two forms from a test given at a single administration allowed us to mimic the typical equating of alternate forms while having the advantage of yielding data from a

single group of examinees that took all of the items on both forms. Because all examinees took both new and reference forms, the two test forms could be directly equated using a single group design, and the result could be used as a criterion to examine the effectiveness of the two linking strategies.

Procedure

Criterion. The criterion represented the true linking of the new form to the reference form. This linking was estimated using a single-group design with those 417 examinees who took both the new and reference forms. As illustrated previously, this single criterion was used to evaluate both the direct anchor linking and the two-stage linking. The schematic of this design is presented in the upper section of Figure 2.

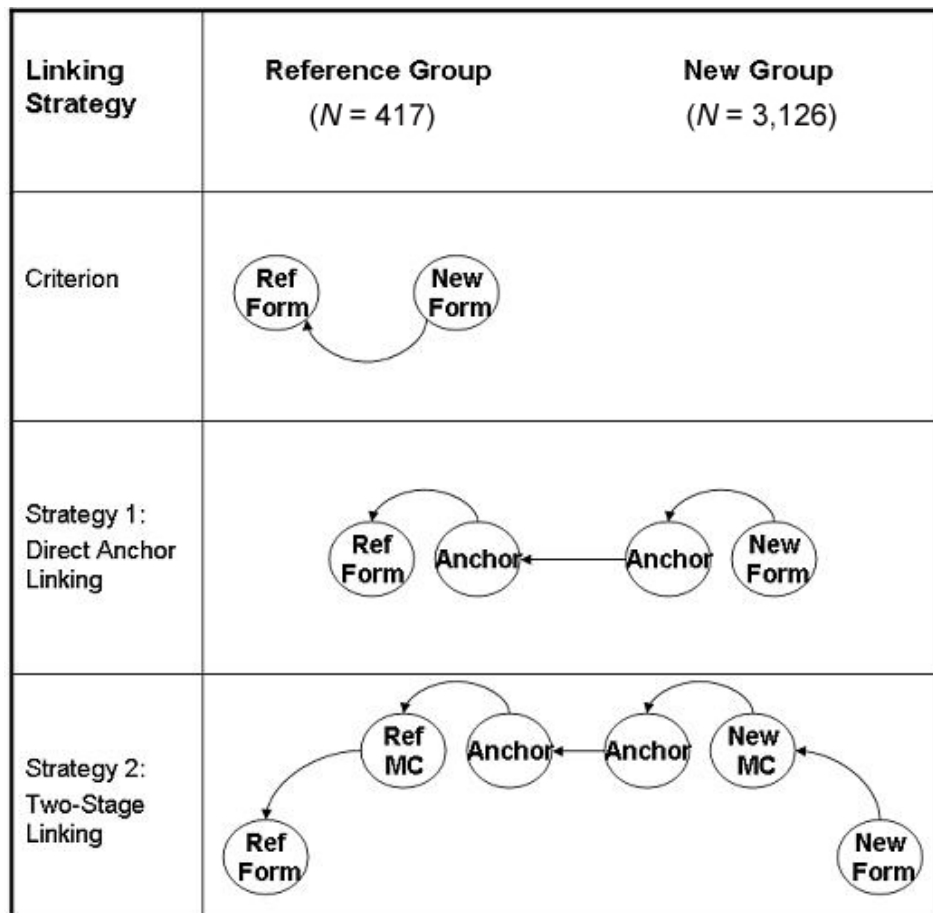


Figure 2. Schematic of the criterion and linking designs examined in this study.

Note. MC = multiple choice.

To estimate the criterion function, total scores on the new form were equated to total scores on the reference form by setting means and standard deviations equal in a single group design. This criterion was used for the linear equating methods. The data were also pre-smoothed using loglinear methods, and a direct equipercentile link was established to produce a nonlinear criterion to use for the nonlinear linking methods.²

Equating methods. Two general linking methods were used: (1) chained linking and (2) post-stratification linking. For chained linking, the chained linear and chained equipercentile methods were used. The Tucker method was selected as the linear version of post-stratification linking. For nonlinear post-stratification, frequency estimation equipercentile linking was performed. These methods and their (untestable) assumptions are described in detail elsewhere (Kolen & Brennan, 2004; Livingston, 2004; von Davier & Kong, 2005).

Equating strategies. Two linking strategies were used in the context of the NEAT design: (1) traditional linking of new to reference form using an MC-only anchor and (2) a two-stage process in which the new MC portion was equated (using chained equating) to the reference MC portion using an MC-only anchor, and then the total composite scores were directly scaled in a single-group design to the MC scores. In practice, the reference MC score was scaled to the reference composite score, whereas the new composite score was scaled to the new MC score. In this way the linked new composite score would be on the scale of the reference composite score (see Figure 1C). The 417 examinees served as the reference population, and the 3,126 examinees served as the new form population. The lower panels of Figure 2 present the schematics of the two linking strategies.

Evaluation. The new form equated raw scores obtained using each linking method under each strategy were compared with the criterion. The differences among the conversions were quantified using the weighted Root Mean Squared Difference (RMSD),

$$RMSD = \sqrt{\sum_{i=0}^{48} w_i [\hat{e}_i(x_i) - e_i(x_i)]^2}, \quad (3)$$

where i represents a raw score point, $\hat{e}_i(x_i)$ is the equated scores of an equating method in a design at raw score x , $e_i(x_i)$ is the criterion equating function at raw score x , and w_i is the relative proportion of the new form examinees at each score point.

Furthermore, a total of 500 bootstrap samples (i.e., 500 replications) were obtained using a resampling technique (random sampling with replacement using the SAS SURVEYSELECT procedure). In each replication, examinees were randomly drawn with replacement from each reference and new form group until bootstrap samples consisted of the exactly same numbers of examinees as in the actual reference ($N = 417$) and new ($N = 3,126$) form groups. Then the new form scores were equated to the reference form for those 500 samples using both the direct anchor and two-stage strategies with the chained and post-stratification linking methods.

In this case, equating bias was defined as the mean difference between chained or post-stratification linking and the criterion linking over 500 replications:

$$Bias_i = \bar{d}_i = \frac{\sum_{j=1}^J [\hat{e}_j(x_i) - e(x_i)]}{J}, \quad (4)$$

where j is a replication, J is the total number of replications (500), $\hat{e}_j(x)$ denotes the raw score equivalent calculated from the chained or post-stratification equating method in sample j , and d_i is the difference between $\hat{e}_j(x_i)$ and $e(x_i)$. The standard deviation of these differences at each score point over 500 replications was used as a measure of the conditional standard error of equating (CSEE) or error due to sampling variability:

$$CSEE_i = s(d_i) = \sqrt{Var_j [\hat{e}_j(x_i) - e(x_i)]} = \sqrt{Var_j [\hat{e}_j(x_i)]}. \quad (5)$$

The sum of squared bias and squared CSEE was considered an indication of total equating error variance at each score point, and the square root of this value defined the conditional Root Mean Squared Error (RMSE) index.

$$RMSE_i = \sqrt{\bar{d}_i^2 + s(d_i)^2}. \quad (6)$$

For overall summary measures, we computed the weighted average root mean squared bias, the weighted average standard error of equating, and the weighted average RMSE across the new form group score distribution.

Results

Descriptive Statistics

Table 2 lists descriptive statistics for the composite tests, MC portions, and MC anchors in the new and reference groups. The table shows that the reference group outperformed the new form group on average on the MC anchor. The standardized mean difference was 0.16. The performance of the reference group on the new and reference forms (recall that this group had scores on both) indicates that the new form was somewhat easier than the reference form. The standardized mean difference between these two scores was 0.07. Table 2 gives the correlations between the MC anchor and other parts of the test for the new and reference forms. As expected, the MC anchor-total MC correlation was higher than the MC anchor-composite correlation. The correlations were fairly comparable across forms, giving an initial indication that (if the assumptions of the research were correct) the tests might be successfully linked using MC-only anchors.

Table 2

Means (Standard Deviations) and Correlations for Examinee Groups Taking New and Reference Forms

Test score (Maximum value)	New group (<i>N</i> = 3,126)	Ref group (<i>N</i> = 417)
New Form X		
Composite (48)	31.32 (6.92)	33.68 (5.97)
Multiple choice (16)	12.04 (2.47)	12.69 (2.16)
Old Form Y		
Composite (48)	—	33.31 (5.96)
Multiple choice (16)	—	12.45 (2.15)
MC-only anchor (8)	5.80 (1.30)	5.98 (1.26)
Composite-MC anchor correlation	.57	.55
Total MC-MC anchor correlation	.83	.83

Note. MC = multiple-choice.

Figure 3 gives the relative frequency polygon for the new form raw score composite in the new form group. The histogram shows a negatively skewed distribution with a mode at 36. By way of reference, the mean for the distribution is 31.3 (see Table 2) and the median is 32. Relatively few examinees score below 20 on the test.

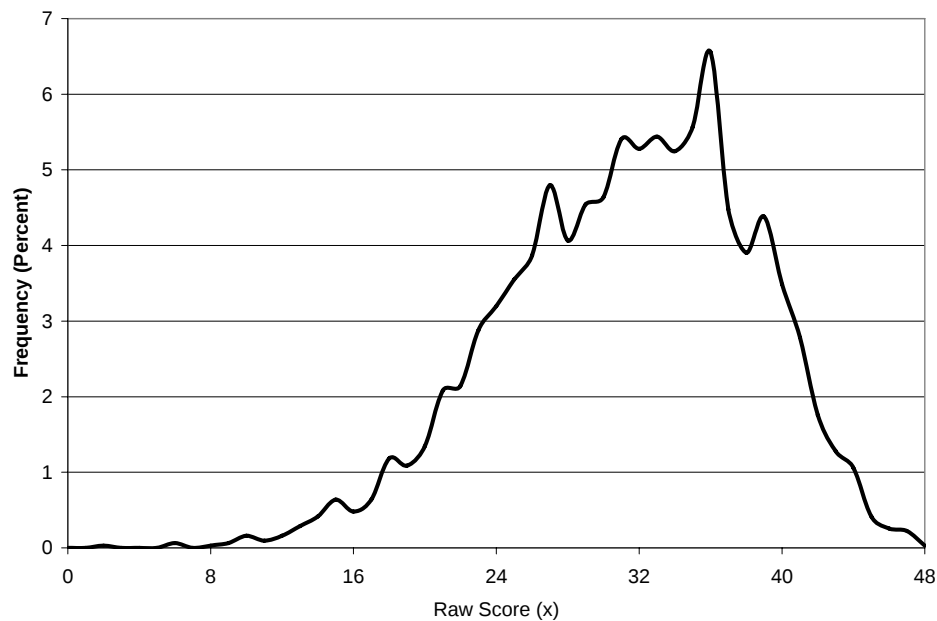


Figure 3. Relative frequency polygon for the new form composite score in the new form group.

Linking Results

Table 3 presents the difference between each linking function and the criterion for the two chained methods and the two post-stratification methods, using the RMSD measure. According to this criterion, the two strategies (direct anchor linking and two-stage) produced equivalent results in the case of chained linking. By comparison, with post-stratification linking methods the two-stage strategy resulted in smaller RMSD measures for both linear and nonlinear methods than the direct linking strategy did.

Figure 4 shows the results for direct anchor linking and for two-stage linking using chained methods, plotted as conditional equated raw score differences from the criterion. The results for chained linear and chained equipercentile linking are shown on the same set of axes for convenience, although the criteria are different for those methods. The figure reflects that the direct anchor and two-stage linking strategies yielded identical results in the linear case. In the

nonlinear case, the differences between the two strategies were negligible in the score range where most examinees were located. The figure also shows that in the nonlinear case, the two-stage strategy did not yield linked scores for the extremes of the score scale, although the direct method did.

Table 3

Summary of Root Mean Squared Difference (RMSD) Between Linking Results and Criterion for Two Linking Strategies with Chained and Post-Stratification Linking Methods

Method	Linking strategy	
	Direct	Two-stage
Chained		
Linear	1.49	1.49
Equipercentile	1.62	1.60
Post-Stratification		
Tucker (linear)	2.00	1.70
Frequency estimation (equipercentile)	2.01	1.82

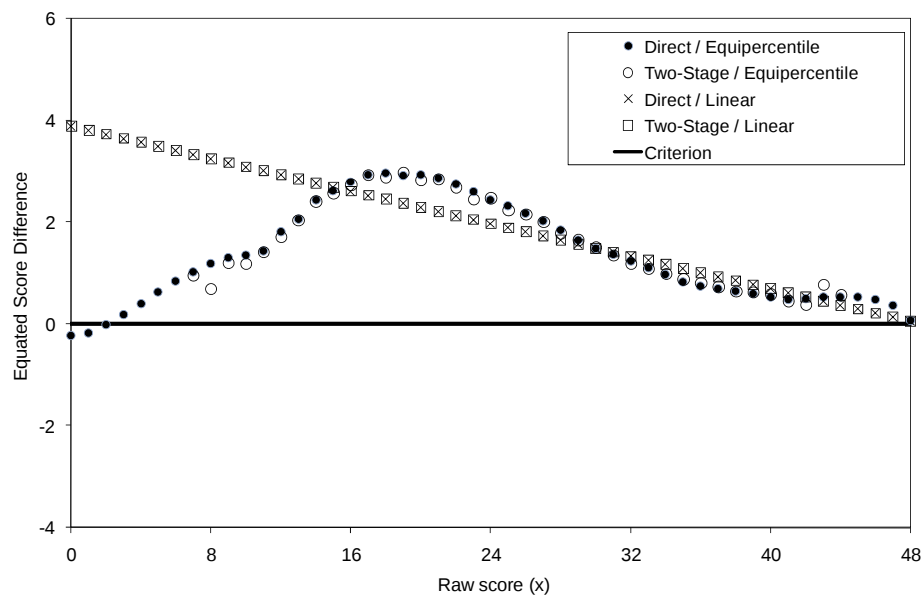


Figure 4. Difference between chained linking and the criterion for the direct anchor and two-stage strategies, chained linear and chained equipercentile linking methods.

Figure 5 shows the corresponding results for the post-stratification linking methods. Results for both Tucker and frequency estimation equipercentile are shown, plotted as conditional equated raw score differences from the criterion. In many places the two-stage equipercentile results are closer to the criterion than are the results of the direct strategy. This advantage is most apparent below a raw score of 20 points. Recall from the histogram in Figure 3 that this score range contains few examinees. Thus the differences between the direct anchor and two-stage results below a raw score of 20 points have little impact on the overall (weighted) RMSD differences. Rather, the equated score differences between the direct anchor and two-stage strategies occurring at score points greater than 20 produce the RMSD differences listed in Table 3. The Tucker results are also somewhat more accurate with the two-stage method than with the direct strategy, especially in the lower end of the distribution.

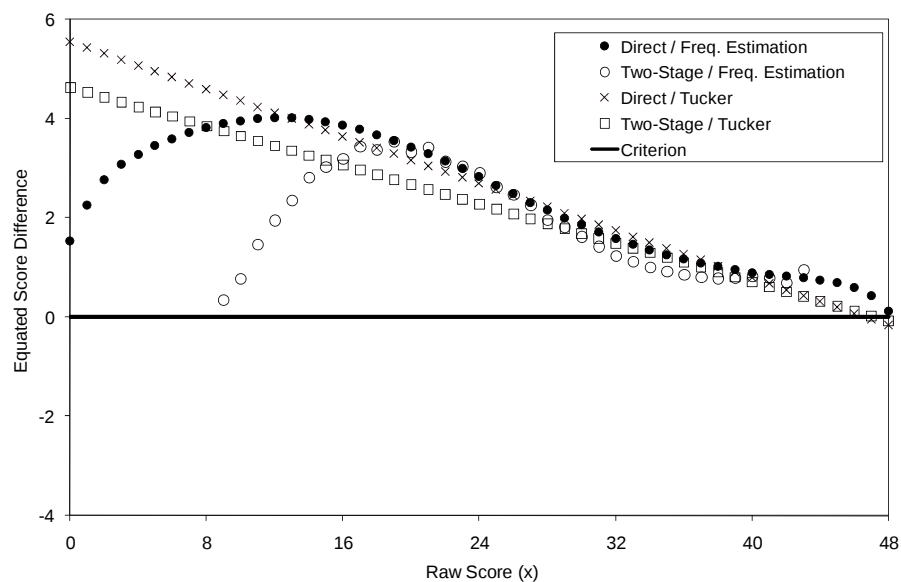


Figure 5. Difference between post-stratification linking and the criterion for the direct anchor and two-stage strategies, Tucker and frequency estimation equipercentile linking methods.

Bootstrapping Results

Table 4 summarizes the deviance measures—root mean squared bias, equating error, and root mean squared error—for the two linking strategies with the chained linking methods. The table shows a substantial amount of bias for both linking strategies, on the order of 1½ score points. As one might expect, the equating error is slightly smaller for the linear than for the nonlinear linking models. There is nothing, however, to distinguish the direct anchor linking from the two-stage linking in the chained case. This fact is illustrated as well in Figures 6 and 7, which show the conditional bias and CSEE for the linear and nonlinear linking procedures, respectively. In Figure 6, only one set of error bands is visible, because the results for the direct anchor and two-stage strategies lie directly on top of one another. For the nonlinear results shown in Figure 7, the direct anchor and two-stage results diverge in the lower part of the score scale, a region in which there are few people (see Figure 3).

Table 4

Summary of Bootstrapped Weighted Average Root Mean Squared Bias, Equating Error, and Root Mean Squared Error (RMSE) for the Two Linking Strategies and Two Methods

Linking method/Strategy	Deviance measure		
	Bias	Equating error	RMSE
Chained linear			
Direct : chained linking	1.51	0.43	1.57
Two-stage: equating and scaling	1.51	0.43	1.57
Chained equipercentile			
Direct: chained linking	1.60	0.63	1.72
Two-stage: equating and scaling	1.59	0.61	1.70

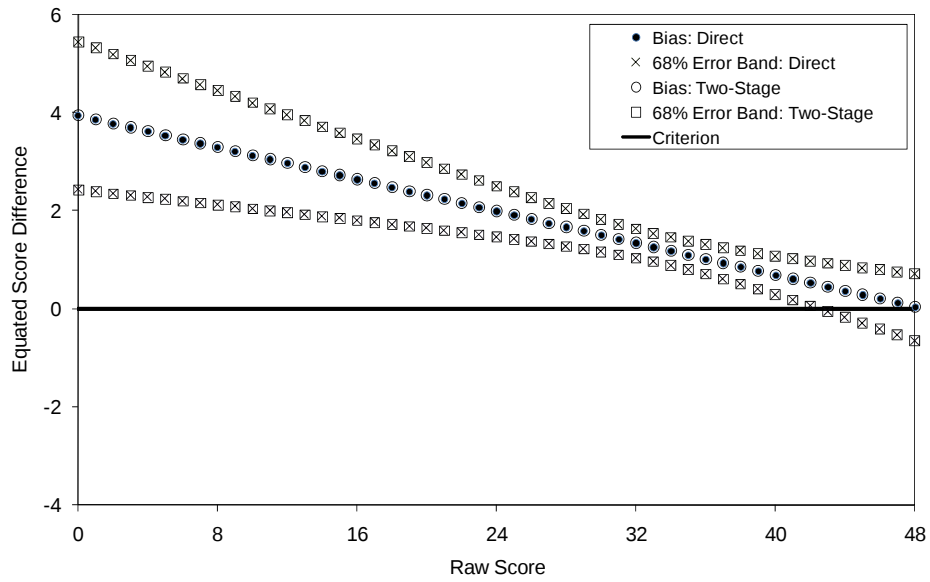


Figure 6. Conditional bias (shown as the average difference of linking results from the criterion) for the direct anchor and two-stage strategies, chained linear case. Linking error bands are also shown.

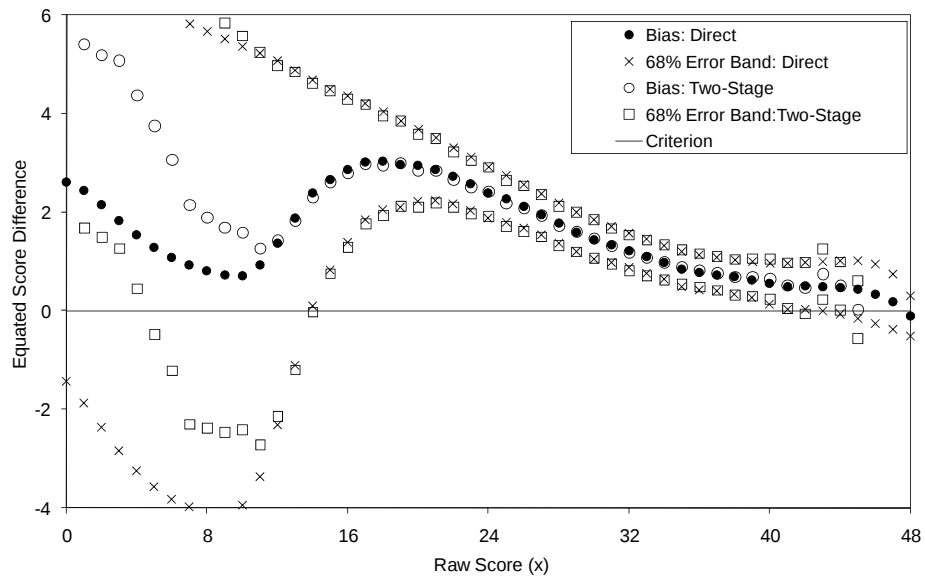


Figure 7. Conditional bias (shown as the average difference of linking results from the criterion) for the direct anchor and two-stage strategies, chained equipercentile case. Linking error bands are also shown.

Table 5 summarizes the deviance measures for the two linking strategies when post-stratification linking methods are used. The table portrays slightly larger bias than with chained methods. The direct strategies result in somewhat larger bias than the two-stage strategies. This is true for both linear and nonlinear linking methods. Figures 8 and 9 illustrate the differences between the two strategies for the Tucker and frequency estimation equipercntile methods, respectively. If we examine Figure 8, we see that the bias curve for the two-stage strategy lies completely within the equating error band associated with the direct strategy. In other words, the reduction in bias obtained with the two-stage strategy is smaller than 1 CSEE. Thus, while the two-stage strategy reduces bias, the reduction is not substantial.

Table 5

Summary of Bootstrapped Weighted Average Root Mean Squared Bias, Equating Error, and Root Mean Squared Error (RMSE) for Post-Stratification Methods

Linking method/Linking strategy	Deviance measure		
	Bias	Equating Error	RMSE
Tucker			
Direct: anchor linking	1.99	.36	2.03
Two-stage: equating and scaling	1.70	.38	1.74
Frequency estimation			
Direct: anchor linking	1.98	.53	2.05
Two-stage: equating and scaling	1.78	.55	1.86

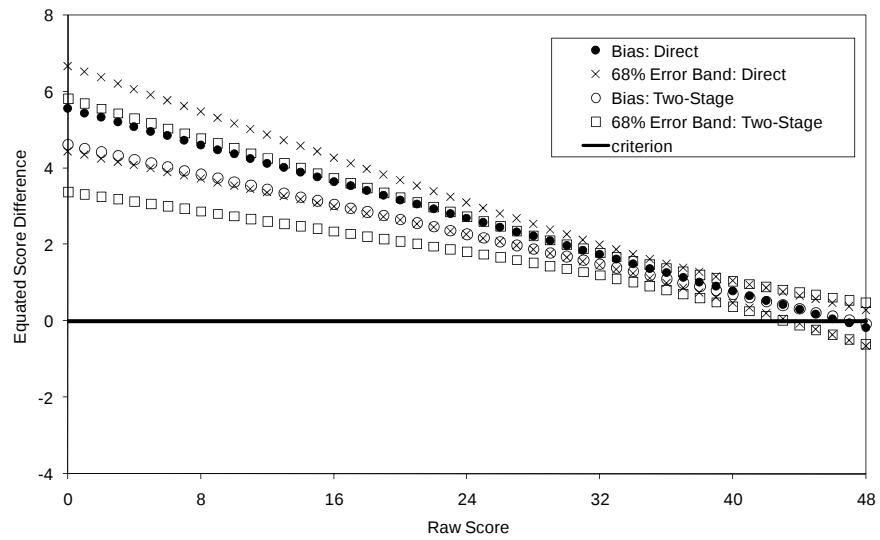


Figure 8. Conditional bias (shown as the average difference of linking results from the criterion) for the direct anchor and two-stage strategies, Tucker case. Linking error bands are also shown.

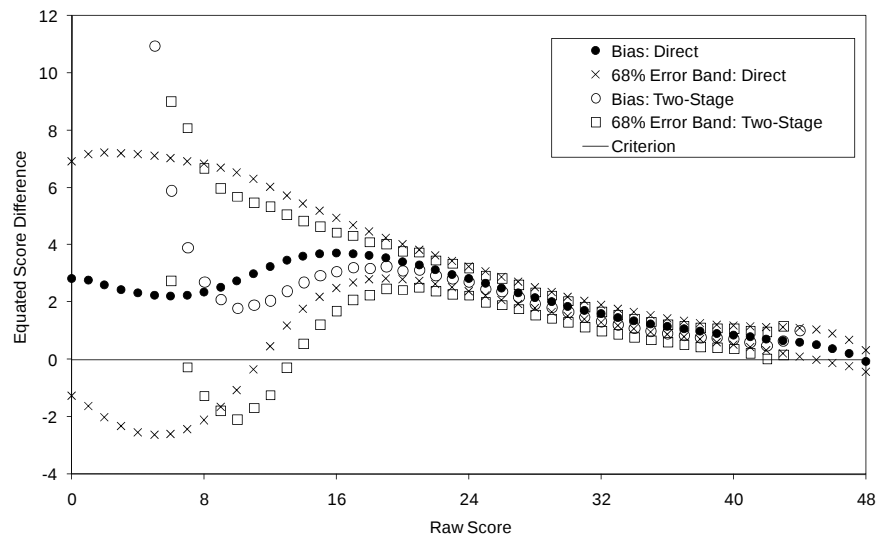


Figure 9. Conditional bias (shown as the average difference of linking results from the criterion) for the direct anchor and two-stage strategies, frequency estimation equipercentile case. Linking error bands are also shown.

Regression Results

The results of this study indicate the presence of substantial bias in equating. One of the suppositions underlying this study was that, even if the MC and the CR portions of the test did not measure the same construct, it might still be possible to use an MC-only anchor to link the reference and new composites, so long as the relationship between the MC portion and the composite were consistent across the reference and new form groups. The similarity of the MC-total correlations across the two groups gave an initial indication that the relationship might be consistent. To examine the consistency of the relationship across the two groups in more detail, the MC and composite scores were first placed on the reference form scale using score conversions obtained through direct scaling in the reference population (recall that this group had scores on all measures). Next, the composite score was regressed on the total MC score in each group separately, using a cubic regression function. Figure 10 displays the results.

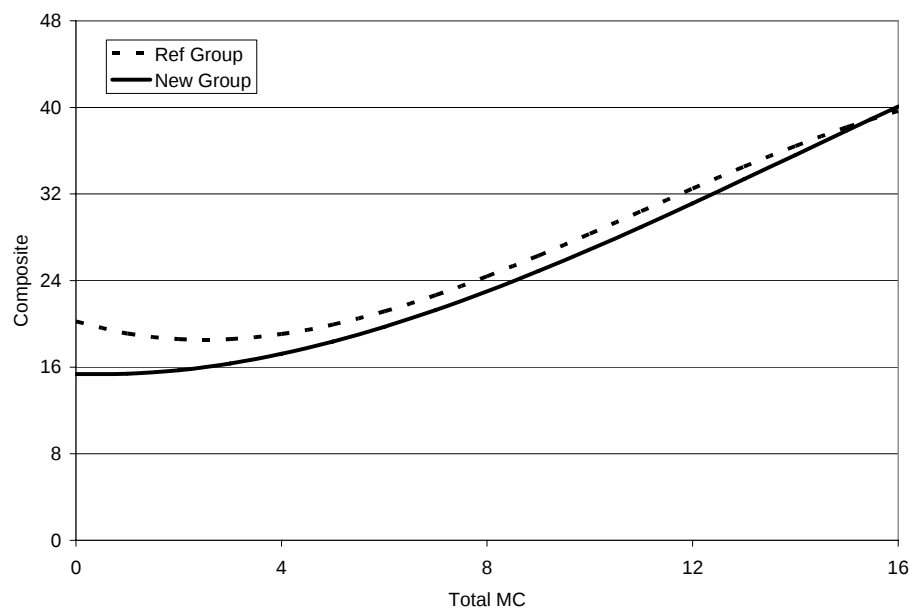


Figure 10. Cubic regression functions relating composite score to total multiple-choice score for the reference and new groups.

Note. Scores are placed on the scale of the reference form.

Figure 10 reveals a substantial difference in the predicted conditional total score means across the two groups. At the lower end of the MC score range this difference is almost 5 score points. Across most of the score range the difference is greater than 1 score point. This difference in the MC-composite score relationship across the new and reference groups could easily result in the substantial linking bias seen here.

Discussion

This study was motivated by the renewed and increased interest in mixed format tests containing both MC and CR items. Informal surveys of the testing industry, coupled with results of our and others' research studies, have convinced us that the linking methods used for mixed-format tests often have not achieved the desired results. We see the problem as twofold. First, there must be some adjustment for differences in rater severity whenever CR items are an integral part of the linking method (i.e., whenever CR items appear in the anchor or whenever equivalent groups designs are used; see Kim et al., 2008, 2010). Second, any mechanism used to adjust for group differences in ability must adjust accurately along whatever dimensions the tests to be linked measure.

In many testing settings, perhaps especially in the K-12 arena, the reuse of CR items (either in an anchor set or in an intact reprinted form) is difficult if not impossible. For that reason, this study examined the possibility of using an MC-only anchor. The study examined two strategies for accomplishing this. One involved using the MC anchor to link composite scores across the new and reference forms. Insofar as the MC and CR portions of the tests measure the same construct (i.e., the test is unidimensional), using MC items alone in the anchor should suffice.

The other strategy used the MC anchor to link the new and reference form MC portions, and then to link the composite and MC scores for each form. We reasoned that the MC anchor would successfully equate the MC portions of the tests. Further, so long as the CR portions projected onto the MC portions in the same manner across groups, the composite scores would be placed on the same scale with one another by scaling the composite scores to the MC scores. We posited that this necessary condition would be met whenever the CR-MC regression was constant across groups.

We demonstrated both analytically and numerically that, while differing in their motivation, the two strategies estimate the same linking function. This fact alone is significant. It

reinforces the notion that what distinguishes equating from scaling is not the particular linking method or strategy used. Rather, it is the nature of the data themselves. Although this point may appear obvious, it often can be overlooked in practice. This oversight can lead to erroneous claims about the scores resulting from such linking.

This research showed that the two linking strategies yielded nearly identical results when chained methods were used. With post-stratification methods, however, the two-stage strategy yielded slightly less biased results. In the present study the very short MC anchor displayed relatively low correlation with the composite score. However, the correlation with the total MC score was much more substantial, as evidenced in Table 2. We would expect that post-stratification methods such as Tucker linear linking or frequency estimation equipercentile linking, which make use of the bivariate moment in computations, would yield more satisfactory results when linking the MC scores as compared with the composite scores, by virtue of the stronger correlation. We saw such an advantage for the two-stage strategy in this instance.

Nonetheless, linking performed on the data set in this study, even with the two-stage strategy, yielded quite biased results. In this sense the results agree with those of previous researchers (Kim & Kolen, 2006; Li, Lissitz, & Yang, 1999). It could be that the linking methods will be successful only if the anchor and the composite score measure exactly the same constructs. Alternatively, it could be, as asserted in this paper, that the composite scores can be successfully linked even if the anchor does not measure exactly the same construct as the composite, in a situation where the relationship between the anchor and the total composite score remains constant across groups.

The similarity of the MC-total correlations gave an initial indication that the relationship was in fact constant across groups. However, the cubic regression results indicated that the MC-composite relationship varied cross the new and reference groups. The separation of the two regression curves in Figure 10 coincided very closely with the conditional linking bias evidenced in Figures 6 through 9.

This finding is not surprising. One of the requirements of equating is that the results be invariant across different subpopulations. This research demonstrated both analytically and with data that in the case of a mixed format test with an MC anchor, the linking procedure can be broken down into an equating problem (MC to MC) and a scaling problem (composite to MC for the new form and MC to composite for the old form). For the composite-to-composite linking to be

invariant across groups, both the equating and the scaling steps must be invariant. We saw with the regressions that the scaling could not in this case be invariant across groups; hence the failure to achieve an equating as evidenced by the large bias. Nonetheless, the results here are equivocal; we cannot conclude from this evidence alone that invariance of the MC-composite relationship across groups (even in the presence of multidimensionality) will allow successful linking. A simulation study examining this possibility would shed more light on the issue.

An interesting finding emerged in the nonlinear case. When the chained and frequency estimation equipercentile methods were applied directly to the composite scores, the resulting new-to-reference form score conversion table covered all possible raw score points on the new form. When the two-stage strategy was employed, however, the score scale was truncated at the top and the bottom, such that no converted scores resulted for the extremes of the new form score scale. This truncation is most likely a result of the concatenation of the linking and scaling results, combined with the sparseness of the data in the extremes of the distributions. Further investigation is warranted to determine the root cause of the phenomenon. However, given the similarity of the results from the two linking strategies, it might appear most prudent to apply the direct anchor strategy as opposed to the two-stage strategy.

A practitioner may still prefer the two-stage method for logistic reasons. Whenever a test contains CR items, test linking is delayed while CR items are scored. Using a two-stage strategy would allow for partial linking of test scores before the CR item scores are available. Then the linking problem, once CR data were available for the new form, would reduce to a quickly implemented single-group procedure. Especially when multiple forms must be equated in a restricted time frame, a two-stage strategy could allow for expedited production of test score conversions once CR scores become available.

As with most research, this study answered some questions, left some unanswered, and revealed still more. The study made explicit the two-stage nature of linking mixed format tests with MC-only anchors. Such an understanding helped to illuminate the possible conditions under which such linking could be successful. The evidence showed that when the regression of CR scores onto MC scores was not constant across the new and reference groups, biased linking resulted. More research must determine if the invariance of the MC-CR regression is a sufficient condition for successful linking or merely a necessary one.

This study negated the belief that a direct linking strategy would yield more accurate results than a two-stage strategy. Indeed, when post-stratification linking methods are advisable, the two-stage strategy would appear to be preferable to a direct one, in that the two-stage strategy yields somewhat smaller bias. The generalizability of this finding, as well as conditions under which it can be used to greatest advantage, have yet to be determined.

This research focused on observed-score linking. Many testing programs that use mixed-format tests, whether by choice, necessity, or contract, utilize item response theory (IRT) calibration methods. Such programs face the same challenges as those mentioned here, in addition to the complication of attempting to fit a unidimensional model to multidimensional data. Observed-score linking methods are not restricted to unidimensional situations in the sense that IRT models are, and thus they may be applicable to a wider variety of situations. Still, some of the findings from the current study may be applicable to testing programs that use IRT methods. Such a possibility provides another avenue for future investigation.

References

- Baghi, H., Bent, P., DeLain, M., & Hennings, S. (1995, April). *A comparison of the results from two equatings for performance-based student assessments*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
- Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement*, 28, 77-92.
- Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9-49). New York, NY: Academic Press.
- Ercikan, K., Schwarz, R., Julian, M. W., Burket, G. R., Weber, M. W., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response test item type. *Journal of Educational Measurement*, 35, 137-154.
- Fitzpatrick, A. R., Ercikan, K., Yen, W. M., & Ferrara, S. (1998). The consistency between raters scoring in different test years. *Applied Measurement in Education*, 11, 195-208.
- Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187-220). Westport, CT: Praeger.
- Kim, S., & Kolen, M. J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Measurement in Education*, 19, 357-381.
- Kim, S., Walker, M. E., & McHale, F. (2008). *Equating of mixed-format tests in large scale assessments* (ETS Research Rep. No. RR-08-26). Princeton, NJ: ETS.
- Kim, S., Walker, M. E., & McHale, F. (2010). Comparisons among designs for equating mixed-format tests in large-scale assessments. *Journal of Educational Measurement*, 47, 36-53.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer.
- Li, Y. H., Lissitz, R. W., & Yang, Y. N. (1999, April). *Estimating IRT equating coefficients for tests with polytomously and dichotomously scored items*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal Canada.
- Livingston, S. A. (2004). *Equating test scores (without IRT)*. Princeton, NJ: ETS.
- Muraki, E., Hombo, C. M., & Lee, Y.-W. (2000). Equating and linking of performance assessments. *Applied Psychological Measurement*, 24, 325-337.

- Sykes, R. C., Hou, L., Hanson, B., & Wang, Z. (2002, April). *Multidimensionality and the equating of a mixed-format math examination*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Tate, R. L. (1999). A cautionary note on IRT-based linking of tests with polytomous items. *Journal of Educational Measurement*, 36, 336–346.
- Tate, R. L. (2000). Performance of a proposed method for the linking of mixed format tests with constructed response and multiple choice items. *Journal of Educational Measurement*, 37, 329–346.
- von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York, NY: Springer.
- von Davier, A. A., & Kong, N. (2005). A unified approach to linear equating for the non-equivalent group design. *Journal of Educational and Behavioral Statistics*, 30, 313-342.

Notes

¹ This is the typical situation for a test with CR items. Each administration involves a different set of raters. The advantage of the data in the current study was the availability of scores for the original reference group from both sets of raters.

² Some may argue that there should be only one criterion, linear or nonlinear. This viewpoint appears reasonable. However, the linear criterion may also be viewed as the linear part in an expansion of the equipercentile function (von Davier, Holland, & Thayer, 2004). If the true criterion function is indeed linear, then both the linear and equipercentile methods will yield the same answer. If the true criterion function is nonlinear, then the linear methods will be estimating the linear part of the equipercentile function, which is correctly represented by the linear criterion.